DOCUMENT RESUME

ED 193 731 CS 503 105

AUTHOR Lewis, James J.; Larsen, Janet K.
TITLE Increasing Inter-Judge Agreement in Forensic Events.
PUB DATE May 80
NOTE 20p.: Paper presented at the Annual Meeting of the International Communication Association (30th, Acapulco, Mexico, May 18-23, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Evaluation: Evaluation Criteria; Evaluation Methods; *Individual Differences; *Judges; *Oral Interpretation; Public Speaking; Secondary Education; Speech Communication; *Training

ABSTRACT

A study was designed to see how judge experience and training sessions influenced interjudge agreement during high school speech contests. Three groups of six judges each viewed a videotape of five poetry readings, gave the performers ratings and rankings, engaged in 30-minute group training sessions, viewed a second videotape of five poetry readings, and rated and ranked the second set of performers. The groups of judges ranged from inexperienced through moderately experienced to highly experienced as oral interpretation judges. Correlations from rankings revealed that training increased the agreement among judges, with agreement after training sessions being quite high; and that training the experienced judges caused the greatest increases in agreement (the more experience judges had, the less they agreed). Experienced judges seemed to use different criteria from one another when discriminating between readers of similar ability, but consistently agreed regarding larger differences of ability between readers. Inexperienced judges seemed to apply one common standard to similar readers, but had no ability to determine overall quality of the group. Training did seem to give inexperienced judges a yardstick of quality. These results suggested ways of maximizing agreement between judges, depending on the number of groups of speakers to be judged, the experience of the judges, training opportunities, and whether each judge evaluated all speakers. (RI)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED193731

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

INCREASING INTER-JUDGE AGREEMENT IN FORENSIC EVENTS

By

James J. Lewis
Janet K. Larsen

DEPARTMENT OF COMMUNICATION
UNIVERSITY OF SOUTH DAKOTA
VERMILLION, SOUTH DAKOTA 57069

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

James J. Lewis

Janet K. Larsen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the International Communication Association Convention, Acapulco, Mexico

May 1980

CS503105

Contemporary individual event speech contests are designed to be educational. All too often, however, the contest experience leaves the participant more confused than educated due to inconsistent judging criticism. As reported by Secord and Thomas, judging decisions are based on the use of different criteria by individual judges and disagreements are often due to the use of different evaluatory standards.[1] While not wishing to make the evaluation of speech events mechanical, the need to improve judge agreement making competition more educational is apparent.[2] The following study analysed the effect of training and judge experience on judge agreement in oral interpretation.

While the need for empirical research in oral interpretation has been recognized for some time, judge agreement in oral interpretation has been researched in only two studies.[3] Larsen analysed judging results from four state sponsored high school contests and found agreement between judge pairs ranged from extreme agreement (r=.99) to absolutely no agreement with approximately one-third of the evaluators showing no similarity to the other judges with whom they were compared. The judges showed more agreement when using rankings than ratings.[4] In the second study, Seedorf conducted four experiments with basic speech students, business students, and oral interpretation students and judges. She reported ". . .little judge with judge agreement. . .".[5] Examining the influence of prior judging, taking an oral

interpretation course, and recieving twenty minutes of criteria explanation, Seedorf found "Agreement among judges does not seem to be affected by either type or quantity of training."[6] Considering the limitations of the Seedorf conclusions, as well as the practical implications of these concerns, further examination of these questions seemed justified.

Therefore, the following study tested the following hypotheses: (1) The amount of judge experience influences agreement between judge evaluations; and (2) Training can increase agreement between judge evaluations.

METHOD

Three groups of six subjects each were selected for participation in the study. The first group, experienced oral interpretation judges, were selected because they had at least a masters degree in speech communication and had judged in no less than ten oral interpretation contests at the regional level or above. The second group, moderately experienced judges, had taken a college oral interpretation class and had judged at local intra-school contests. The third group, the inexperienced judges, consisted of persons who had neither taken any oral interpretation class nor had judging experience.

Two sets of video-tapes were constructed to be shown to the subjects. Each tape set contained five poetry readings by five different readers. These readings were selected by first recording the performance of 16 student readers from

beginning and advanced oral interpretation classes. The five poorest readings, as determined by the experimentor, were eliminated. Of the eleven remaining, the top five were used as the "Better Speaker" set, the middle performance was used in training sessions, and the last five comprised the "Poorer Speaker" set. The order of the readings on the tape was randomly assigned.

The study used a pre-test post-test design, with training serving as the independent variable. The subjects saw one set of tapes, gave the readers both ratings and rankings (the dependent variables), engaged in training, and viewed and gave ratings and rankings to the readers on the other tape set. Rankings were from one to five, with no ties allowed. The rating scale ranged from 1 to 32 and ties were allowed. Counter balancing was used to control for the different tape sets, with half the subjects in each group evaluating the "Better Speaker" set first and the other half evaluating it second.

The three subjects from each group who watched the same tape set together were also trained together. The training lasted about 30 minutes. First, a specific criterion for judging was distributed and discussed. The subjects viewed and evaluated the "training session reader." They next shared their evaluation and discussed differences with each other and the trainer.

RESULTS

Rankings and ratings of the subjects were examined to explore the influence on subject agreement of (1) the subjects' judging experience, and (2) training. Scores were first placed in an analysis of variance model corresponding to the design used, and results of this analysis were used to calculate intraclass correlation coefficients to determine subject agreement.[7]

In addition to variation due to subjects' judging experience and training, the analysis of variance model identified the following sources of variation:

1. The average score given by each subject (this source of variability can also be thought of as the score the judge would give the average speaker, and is thus termed the "anchor point" for that judge),
2. The particular reader being evaluated as compared with the other four readers on the tape set,
3. The difference between the two tape sets, and
4. Random error.

These sources of variation can be viewed several ways to provide intraclass correlations (the ratio of systematic variation to systematic plus error variation).

The sources of variation for intraclass correlations based on rankings, for example, are the particular reader (source #2 above) and random error (source #4). Since the subjects were forced to give the ranks of first through fifth, the median score--given by each subject, to each tape set, before and after training, and regardless of subject experience--had to be "third". No variability exists due to these factors. The analysis of variance results for ranks are summarized in Table 1, and the intraclass correlation coefficients showing agreement between the three subjects viewing the tape are given.

TABLE 1 About Here

Rating scores, however, contain all the sources of variation listed above. The intraclass correlations depend on which of these sources of variation are defined as systematic and which are defined as error. Three different definitions of these systematic and error terms, resulting in three different sets of coefficients, are discussed below. Analysis of variance results for ratings are summarized in Table 2, using both random error and random error plus anchor point differences (source #1) as the error term.

TABLE 2 About Here

Scores were then pooled across the two tape sets ("Better Speaker" compared to "Poorer Speaker" tape sets

did not significantly influence the extent of subject agreement) and intraclass correlations based on experience and training were computed. These correlations are given in Table 3.

TABLE 3 About Here

Correlations for rankings revealed two obvious patterns (see Table 3). First, training increased the agreement among subjects. Agreement after training was quite high for all groups (more than .6). Second, the more experience subjects had, the less they agreed. Experienced subjects did not agree before training (these correlations were not significantly different from chance). Training, in this instance, caused the greatest increase in agreement when the subjects were experienced.

The same pattern appeared using model 1 to analyze ratings (see Table 3). This model used, for the intraclass correlation coefficients, exactly the same sources of variance as were used for rankings; random error (source #4), and the particular reader (source #2). These are the only sources of variation present when judges are evaluating one set of speakers of similar ability. Under these conditions, training was again extremely effective for increasing agreement, especially among experienced subjects. Without training, inexperienced subjects showed more agreement than experienced subjects.

Other sources of variation resulting from different

situations necessitated different models for analysis and changed the pattern of results. Model 2 was the same as model 1 except that it also included variation between the "Better Speaker" and "Poorer Speaker" tape sets (source #3) as systematic. This model represents a situation where judges evaluate more than one section of speakers, and the sections are of differing ability levels. The intraclass correlations again show that training increased the amount of agreement, but agreement was similar for all experience levels. Based on the results of models 1 and 2, experienced subjects apparently used criteria different from one another when discriminating between readers of similar ability, but consistently agreed regarding larger differences of ability between readers. Inexperienced subjects seemed to apply one common standard to similar readers, but had no ability to determine to overall quality of the group (there was little difference in the correlations for inexperienced subjects on the pretest between models 1 and 2). However, training seemed to give inexperienced subjects a yardstick of quality, as this group showed the largest post-test increase when comparing models 1 and 2.

Model 3 employed the same sources of systematic variation as did model 2, but also considered differences in anchor points (source #1) as part of error variation. This model would apply when each judge does not evaluate all speakers in an event. If each judge evaluated all speakers,

everyone would be affected equally by a judge giving consistently low scores (as in model 2). When not all speakers are judged by the low evaluator, then anchor point differences influence the final position of the speakers.

Regarding experience, intraclass correlations for model 3 showed the reverse of the model 1 pattern (see Table 3). In this model, experienced subjects demonstrated more agreement than inexperienced ones. While training increased agreement, inexperienced subjects improved markedly while experienced subjects improved little. This indicated that experienced subjects had similar anchor points, while the anchor points of inexperienced subjects were dissimilar. Training effectively established similar anchor points for inexperienced subjects.

DISCUSSION

Results of this study suggest applications for evaluation of speech activities. In situations where agreement between judges is important (learning theory suggests it may be quite important) and speakers are of relatively similar ability, specific steps can be taken to maximize agreement. The following suggestions must, however, be considered in light of the limitations of the study. Three major limitations need to be considered.

The present study used only poetry selections. The degree to which judging dynamics may be similar with different genrés, or other forensic activities, has not been an-

swered. The dynamics may also be more or less similar depending on the experience level of the judges.

The present study was also limited by regional sampling. Major regional differences in speaking and/or judging would restrict the generality of results.

A final limitation resulted from the design. Readers were so grouped that those with similar ability levels appeared on the same tape. If speakers were not grouped homogeneously, the same pattern of results might not have occurred, particularly with inexperienced judges.

Remaining cognizant of the limitations stated above, suggestions for maximizing agreement depend on the following conditions: whether one or more groups of speakers are to be judged, the experience of the judges, the possibility of training, and if each judge evaluates all speakers.

When only one group of speakers is being judged, ranks should be used instead of ratings. Without training, inexperienced judges provide the greatest amount of agreement. Training will increase the amount of agreement for all, and experienced judges should not be used without training.

If several speaker groups with differing ability are to be judged, and each judge evaluates all speakers, then rating scales should be used. Experienced and inexperienced judges have similar levels of agreement, and while training helps, it is especially valuable for the inexperienced judge.

When several groups of speakers with differing abil-

ities are judged but not all speakers are evaluated by each judge, rating scales are again preferred. Experienced judges provide the greater agreement, and inexperienced judges should not be used without prior training.

If training is not possible, the type of evaluation scale should depend on the experience of the judges. Experienced judges should be given rating scales, and inexperienced judges should assign ranks.

When the possibility of training exists, and maximum agreement is desired, training should be used. Training is particularly important when inexperienced judges are ranking several groups but not evaluating every speaker. Training seems to do two things for the inexperienced judge; first, it provides a standard of quality independent of the particular speaker group, and second, it provides a common anchor point. Training for experienced judges evaluating speakers of similar ability is also important, but for a different reason. For experienced judges, training can establish a common criteria for consistently distinguishing between small differences in ability.

The present study indicates that practical results can be achieved by applying emperical research methods to the judging of forensic activities. The suggestions of this study should be verified and expanded by using less homogeneous groups of speakers, different judges, and a variety of forensic events. Methods of effective and

practical training provides a second area for investigation. Finally, future research should determine the kinds of evaluation that maximize participant learning. Factors other than agreement need to be explored. Differing factors may become important for different speaker subgroups. By better understanding the judging process, studies such as these can increase the educational value of the contest experience.

FOOTNOTES

[1]Arthur E. Secord and Ruth H. Thomas, "Speech in the Extra Curricular Program," The Role of Speech in the Secondary School, Bulletin of the National Association of Secondary School Principals, XXIX (November, 1945), p. 6.

[2]See Virginia E. Martin, "Evaluation of High School Declamation in Selected Midwestern States," Thesis Univ. of So. Dakota 1955; and Janet K. Larsen, "An Experimental Study to Increase Judge Agreement of Oral Interpretation Performances," Thesis Univ. of So. Dakota 1977.

[3]The need was well stated by Beverly Whitaker, "Research Directions in the Performance of Literature," Speech Monographs, XL (August, 1973), p. 241. Studies exploring agreement were Evelyn Seedorf, "An experimental Study of the Amount of Agreement Among Judges in Evaluation of Oral Interpretation," Diss. Univ. of Wisconsin 1947; and Janet K. Larsen, "Judge Agreement in Oral Interpretation," South Dakota Speech Communication Association Speech Journal, IV (Spring, 1977), pp. 14-16.

[4]Larsen, footnote 3, p. 16.

[5]Seedorf, p. 161.

[6]Seedorf, p.

[7]This statistic is explained and its application to ranks is justified by Ernest A. Haggard, Intraclass Correla-

tion and the Analysis of Variance, (New York, The Dryden Press, Inc., 1958).

TABLE 1

BETWEEN AND WITHIN GROUP MEAN SQUARE VALUES AND INTRACLASS CORRELATIONS FOR EACH GROUP OF RANKS

Stimulus Tape	Experienced Judges 1st group	Experienced Judges 2nd group	Moderately Experienced Judges 1st group	Moderately Experienced Judges 2nd group	Inexperienced Judges 1st group	Inexperienced Judges 2nd group
Pre Training						
Better Speakers						
MSBG	2.833		4.167		5.833	
MSWG	1.867		1.333		.667	
r	.147		.415		.721**	
Poorer Speakers						
MSBG		3.333		4.833		4.833
MSWG		1.667		1.067		1.067
r		.250		.541*		.541*
Post Training						
Better Speakers						
MSBG		4.500		5.333		5.167
MSWG		1.200		.867		.933
r		.478*		.632**		.602*
Poorer Speakers						
MSBG	6.667		5.833		6.833	
MSWG	.333		.667		.267	
r	.864**		.720**		.891**	

Note. Significance tests are based on F's with 4 and 10 degrees of freedom.

* pH_0 ($r = .00$) $\leq .05$.

** pH_0 ($r = .00$) $\leq .01$.

TABLE 2

BETWEEN AND WITHIN GROUP MEAN SQUARE VALUES AND INTRACLASS CORRELATIONS FOR EACH GROUP OF RATINGS

Stimulus Tape	Experienced Judges 1st group	Experienced Judges 2nd group	Moderately Experienced Judges 1st group	Moderately Experienced Judges 2nd group	Inexperienced Judges 1st group	Inexperienced Judges 2nd group
Pre Training						
Better Speakers						
MSBG	17.833		14.000		66.900	
MSWG	18.033 (16.400)		7.050 (5.800)		10.550 (12.133)	
r	-.004 (.028)		.247 (.320)		.640* (.601)*	
Poorer Speakers						
MSBG		50.433		55.567		84.933
MSWG		25.883 (23.867)		17.254 (16.467)		32.733 (36.200)
r		.240 (.248)		.425 (.442)		.347 (.310)
Post Training						
Better Speakers						
MSBG		32.433		90.733		41.067
MSWG		8.633 (8.400)		14.383 (12.400)		9.867 (9.133)
r		.479 (.488)*		.639* (.678)**		.513* (.538)*
Poorer Speakers						
MSBG	78.900		37.067		89.100	
MSWG	7.100 (16.933)		4.917 (12.867)		20.100 (31.133)	
r	.771** (.550)*		.686** (.385)		.534* (.383)	

Note. Numbers given in parentheses have included variation due to anchor point differences between judges in the error term. This error term is used for computing model 3 only (see text or footnote for Table 3). Significance tests using the parenthesized value are based on F's with 4 and 16 degrees of freedom. Tests using the non-parenthesized values have 4 and 8 d.f.

* pH_0 ($r = .00$) $\leq .05$.

** pH_0 ($r = .00$) $\leq .01$.

TABLE 3

COMPARISON OF INTRACLASS CORRELATIONS
FOR POOLED GROUPS BASED ON DIFFERING MODELS

Data	Training	Experienced Judges	Moderately Experienced Judges	Inexperienced Judges
Rankings				
	Pre	.155	.438*	.598**
	Post	.646**	.646**	.724**
Ratings: Model 1				
	Pre	.131	.362*	.435*
	Post	.655**	.652**	.508**
Ratings: Model 2				
	Pre	.516**	.504**	.476*
	Post	.772**	.855**	.831**
Ratings: Model 3				
	Pre	.506**	.495**	.396*
	Post	.633**	.786**	.724**

Note. Data for this table were pooled across both stimulus tapes. Different models refer to different variance components that were used for computing the intraclass correlation coefficients. Model 1 used residual variance as error variance, and variability of each speaker compaired with the other four speakers on the same stimulus tape as systematic variance. Model 2 also included variance due to differences between the stimulus tapes as systematic. Model 3 used the same sources of variance as Model 2, and added variability due to anchor point differences of judges to the error variance. Significance tests were based on F's with 9 and 20 degrees of freedom for Rankings and for Model 1, with 9 and 18 d.f. for Model 2, and with 9 and 17 d.f. for Model 3.

* pH_0 ($r = .00$) $\leq .05$.

** pH_0 ($r = .00$) $\leq .01$.